

AD-A083 048

ILLINOIS UNIV AT URBANA-CHAMPAIGN DEPT OF PSYCHOLOGY

F/6 5/10

ANALYZING INFREQUENT EVENTS: ONCE YOU FIND THEM YOUR TROUBLES B--ETC(U)

APR 80 C L HULIN D M ROUSSEAU

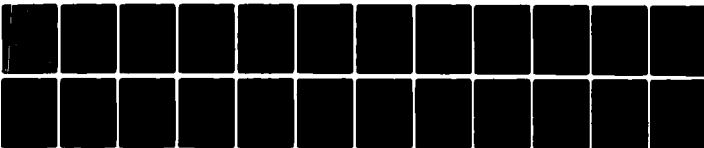
N00014-75-C-0904

UNCLASSIFIED

TR-80-3

NL

171
AD 3
7/8/82



END

DATE

FILED

5-80

DTIC

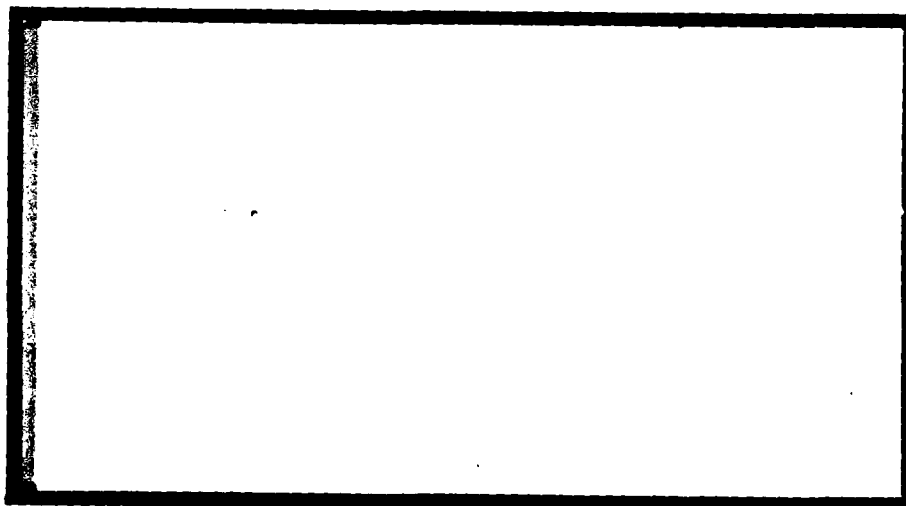
LEVEL II

(12)

UNIVERSITY OF ILLINOIS

Studies of Individuals and
Groups in Complex Organizations

ADA 083048



DDC FILE COPY

Department of Psychology
Urbana - Champaign

DTIC
SELECTED
APR 15 1980
E

80 4 14 047

ANALYZING INFREQUENT EVENTS:

ONCE YOU FIND THEM YOUR TROUBLES BEGIN.

Charles L. Hulin
University of Illinois at Urbana-Champaign

Denise M. Rousseau
United States Navy Postgraduate School

Technical Report 80-3
Apr 1980

Running Head: Analyzing Infrequent Events

Prepared with the support of the Organizational Effectiveness Research
Programs, Office of Naval Research, Contract N00014-75-C-0904 NR 170-802.

Reproduction in whole or in part is permitted for any purpose of the United
States Government.

Approved for public release; distribution unlimited.

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM	
1. REPORT NUMBER 80-3	2. GOVT ACCESSION NO.	3. DISTRIBUTION STATEMENT (When Data Entered)	4. SECURITY CLASS. (When Data Entered)
5. TYPE OF REPORT & PERIOD COVERED Analyzing Infrequent Events		6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s) Charles L. Hulin Denise M. Rousseau		8. CONTRACT OR GRANT NUMBER (if any) N000-14-75-C-0904	
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Psychology University of Illinois Champaign, IL 61820		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NA170-802	
11. CONTROLLING OFFICE NAME AND ADDRESS Organizational Effectiveness Research Programs Office Of Naval Research (Code 452)		12. REPORT DATE April, 1980	
13. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Arlington, VA 22217		13. NUMBER OF PAGES 24	
		14. SECURITY CLASS. (of this report) Unclassified	
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE	
15. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.			
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)			
18. SUPPLEMENTARY NOTES			
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) low base rate phenomena, Poisson distribution, infrequent events, aggregation across time, vertical aggregation across groups, latent trait analysis of infrequent events.			
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) < A theoretical/methodological analysis of the study of infrequent or low base rate events and behaviors is presented. Problems of interpretation caused by aggregation across arbitrary time limits imposed by the necessity to obtain variance in dependent measures, aggregation across heterogeneous samples drawn from unspecified populations, studies of rates of occurrences per group, studies of surrogates with less extreme base rates, and studies of post hoc groups defined after the occurrences of the event			

20. (continued)

are all discussed. The problems with each attempt to circumvent extremes of base rates were considered. Available alternatives were presented and discussed.

Accession For	
NTIS GINA&I	<input checked="checked" type="checkbox"/>
DDC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Accession Codes	
Dist	Mail and/or special
A	

If we observe any given person on a randomly chosen day, the chances are overwhelming that the person will arrive at work on time and will perform adequately--neither so well nor so poorly as to merit comment. The probability of any person, selected a priori, having an accident, being a victim of a serious crime, or performing exceptionally on the job is low. These infrequent events and behaviors are considered "low base rate phenomena." Of course, the obverses of the above statements are also true; if we want to predict what a person will do on any given day we can capitalize on the base rate and bet they will do nothing exceptional.

We are interested, both as practitioners and theoreticians, in predicting and accounting for variance in the occurrence of infrequent events. These events attract our attention. They are frequently costly, and often signal the breakdown or imminent breakdown of a system. Others, such as assassinations, suicides, and natural disasters are important because of their signal value and their costs to society.

Still other low base rate events may be useful as surrogates for other phenomena that have even lower base rates but demand study because of their costs when they do occur--absenteeism used as a surrogate for turnover, suicide attempts as surrogates for suicides, local severe pockets of unemployment or recessions used as surrogates for depression. All too often the theoretical relevance of the surrogate to the latent construct is assumed rather than demonstrated.

The very characteristics that make these events interesting to us also make them difficult to study. We may study a group of individuals over a period of time and observe, at worst, no occurrences of events of interest, or at best, little variance in the distribution of the events across individuals. Many a well designed study of turnover or accidents fails to

investigate adequately relations between low base rate events and other variables because of vanishingly small variance in the criterion. In general, the ratio of exceptional to everyday events or behaviors may approach zero.

Investigators faced with very low base rates of the behavior or event they wish to study often adopt a number of strategies designed to relieve the most serious distributional problems of these infrequent events. Some extend the time period during which the behaviors are observed and collapse observations across these longer periods. Others expand their samples from the original to new populations and collapse individuals into one sample. Others study rates of the occurrence of the behavior or event in collectives of individuals. Some redefine their criterion to one less extreme with a more favorable base rate. Still others analyze samples that have been constituted on a post hoc basis--after the events or behaviors have occurred. All of these strategies solve the base rate problem but create others.

Some Brief Distributional Notes

Perhaps the most frequent statistical analysis is to test whether the conditional distribution of x_1 (any interesting event) following (or simultaneous with) another event x_2 , $P(x_1 | x_2)$, is the same as the unconditional distribution of x_1 , $P(x_1)$. In the simplest case, if the probability of observing x_1 is the same whether x_2 occurred or not, the two events are independent, uncorrelated and, x_2 is said not to cause x_1 . If x_1 always occurs following x_2 and never occurs without x_2 , x_1 may be dependent on x_2 , is perfectly correlated with x_2 , and x_2 may be a cause of x_1 .

In the study of low base rate phenomena, analysis of the unconditional distributions of x_1 usually indicate departure from normality because low

base rates mean, literally, a few occurrences must occur within a large groups of individuals. This implies that for many individuals the event never occurs and the resulting distribution will be positively skewed. One does not usually think of testing the shapes of obtained distributions against theoretical distributions (such as the normal, binomial, or Poisson distribution) as a preliminary step in most analyses. However, one candidate for a "null" hypothesis for the distribution of rare events is a theoretical distribution that would be obtained if the events were to occur randomly and independently across individuals where every individual in the sample is equally likely to experience the event, $P_i = P_j$, and future occurrences of the event are independent of past occurrences. Such a process could generate a Poisson or a binomial distribution (Feller, 1950, p. 142-155). Examples of events following Poisson distributions are the frequencies of being kicked to death by a mule in the Prussian army, the arrival times of cars at a toll booth on the New Jersey Turnpike, emission of particles per unit of time in radioactive decay, and the occurrences of bacteria per unit area in infected organisms. Individual occurrences of these events are difficult to predict because they seem to occur randomly and independently with respect to individuals or may be conditionally random for employees within a particular industry performing a particular job. Accidents occur more frequently in some industries but individual occurrences appear independent of individual characteristics.

A frequent example of Poisson distributions in social science research is found in the occurrences of accidents suffered by individuals in a given period of time. If the sample size, N , is large and the probability of an accident, P , is small, then we can define $NP = \lambda = \text{sample mean}$. Under these conditions, the Poisson has an interesting distributional property; the mean

of the distribution is equal to the variance ($\bar{x} = s$). This provides a method of estimating the goodness of fit of the obtained distribution to the theoretical distribution. Where the mean and variance are equal, the obtained distribution demonstrates some critical parameters of the Poisson. If the mean of the distribution, \bar{x} , is a function of the danger of the job, then individual differences would not influence this parameter. If the variance of the obtained distribution is fit by a random process without provision for individual differences, then it might be concluded that further analysis of individual antecedents of the events or behaviors under study, at least for this sample, is fruitless because no process can improve on the assumed underlying random process that makes no provisions for unequal distributions of the tendency to experience the event or engage in the behavior.

Although this is a tempting conclusion, it would be premature. Random assignment of subjects to groups cannot be assumed in most non-experimental, social science research. If individuals with high accident liabilities, individuals who are accident prone, are attracted to jobs with high danger levels while those low in proneness are attracted to jobs with low danger levels (rodeo cowboys versus clerical workers) then the means of the obtained distributions would be influenced by a combination of environmental and individual factors. Non-randomness across time revealed by non zero correlations of accidents suffered in two time periods could also invalidate the conclusion.

An examination of the formula specifying the frequencies of the occurrences of exactly 0, 1, 2, ..., k occurrences of an event suggests that with small values of P, the overwhelming majority of the individuals in any sample will have 0 occurrences, a few will have 1, and almost nobody will

have more than 2. In general, $P(n) = \frac{e^{-\lambda} \lambda^n}{n!}$ where n is the number of events and the other terms are as defined above. The resulting distribution will be badly skewed and will have very small variance. Further examination of the equation, however, suggests at least two ways out of the dilemma, one way to increase the number of events without causing them yourself, and at least one tempting alternative research strategy. These are discussed briefly below.

Horizontal Aggregation Across Time

One way to increase P , the likelihood of observing an infrequently occurring event, is to increase the length of time over which observations are made. If we wish to witness a May blizzard in the Midwestern United States or an earthquake in California, we need only wait. In observing an individual's behavior for instances of tardiness or accidents, the greater the time interval over which we make observations, the more likely we are to observe variation in his or her behavior. Increasing P increases NP , λ , and s^2 . Whenever we collapse data gathered over a long period of time we perform a horizontal aggregation. The researcher must choose an appropriate time interval for aggregation whenever this is done. Because time intervals are so important to variability in low base rate phenomena, it may be argued that the longer the time interval the better. However, by increasing the interval over which data are aggregated we encounter: fluctuations in relations among variables, history/maturation confounding, and lack of synchronicity.

Fluctuation in Relations

In any study, variance is needed for relations or effects to be demonstrated. Hence, if motivation, assessed at the beginning of the time period, and subsequent absenteeism are to be related to each other, the

greater the variability in absenteeism, the greater the potential correlation between these variables. Variability in absenteeism is increased as we increase the time over which absence data are collected. However, increasing variance may not consistently lead to a higher correlation. We might expect the correlation between motivation and an aggregated measure of absenteeism will decrease as the time interval is increased beyond an appropriate length. Changes in other variables may alter the relations. If we measure satisfaction in January and correlate it with absenteeism aggregated over one month, two months, three months and so on until the next December, the correlations might first increase because of increases in variance and then begin to decrease with the increasing time interval. Satisfaction changes over time or non-attitudinal factors (e.g. home life, weather, available alternative activities) influence absenteeism but are not constant. How long a time interval is needed to demonstrate a relation between the variables? What is the "appropriate" time interval for a study of absenteeism or any other human behavior? Most theories are mute on this question.

In a study exploring the effects of time on correlations involving aggregated data (absenteeism), Rousseau (1978) found the variance in the aggregated variable did indeed increase as the time interval increased. However, when correlations between satisfaction and the aggregated variable were computed, the correlations first increased for the one, two, and three month intervals and then decreased as the time interval was lengthened to four months. Because researchers often aggregate infrequent events such as absenteeism over a period of as long as a year (e.g. Hackman & Lawler, 1971; Nicholson, Brown & Chadwick-Jones, 1976) they may provide a very time specific estimate of the relation of the aggregated variable and other

measures. In social science research, in fact in nearly all research on living organisms, the time interval chosen over which data are aggregated may seriously affect empirical findings. Few researchers explain their choice of time intervals. Fewer theories address issues of the time periods across which data should be gathered and relations among variables would be expected to hold. Intervals in empirical research seem to be chosen on the basis of periodicity foreign to the event being studied--lengths of grants, academic years, semesters, or economic cycles seem to dominate our choices. In the biological sciences, time intervals are dictated by the life cycles and rhythms of organisms under study. Such natural cycles may exist but seem unknown in most social sciences.

The researcher's dilemma is how to achieve increased variance in the distribution of low base rate phenomena without going beyond the limits of the natural time intervals or cycles. Exceeding these (unknown) limits may well obscure relations that exist between the low base rate variables and their antecedents. Systematically exploring effects of time intervals on relations will generate an empirical basis for solving this problem. Time series analysis or cross lagged correlational studies with many different lags are possible methodological bootstrapping techniques that can be used to address this theoretical problem. Time must be made an explicit component of both theory and empirical research designs.

History/Maturation

A related issue is lawful changes in the individuals observed in the time interval $i+1$ as a result of an occurrence or non-occurrence of an event during the i th interval. Does an absence, for example, during one time period increase or decrease the probability of an absence during subsequent time periods? Does a suicide attempt by an individual influence the

probability of observing a suicide by that same person at a later time? Do accidents inoculate individuals against later accidents? If probabilities of occurrences are not independent across time, we expect gradual changes in the empirical meanings and antecedents of the events at the group level (even though our units of observation and analysis are individuals, our statistics are based on group effects) as a result of more and more members of the group having accidents, periods of absenteeism, or attempting suicides. At the individual level, the change in the meaning of the event would be abrupt following the occurrence of the event for the first, or even the *i*th time. At the group level, these small abrupt changes in individuals cumulate to produce gradual systematic changes in group data. Abrupt discontinuities at the individual level that generate smooth curves at the group level should be the targets of fine grained analyses applied across time at the individual level.

Lack of Synchronicity

Horizontal aggregation across time introduces a third problem: lack of synchronicity. Synchronicity exists when two related variables are measured at the same time and have the same time referents. Variables correlated in longitudinal research often refer to different time periods. For example, when tardiness is correlated with a measure of employee motivation, tardiness may be measured over a month or a year. Motivation is usually measured at a single point and may reflect either a short-term self-perception or a chronic attribute of individuals' personalities. It is doubtful that motivation is bounded by the arbitrary time interval over which tardiness is measured. In this example, synchronicity may not exist.

When correlations are used to test models in which causal relations are hypothesized, lack of synchronicity poses an interpretational problem

(Kenny, 1975). When researchers measure employee attitudes or perceptions they often assume that attitudes and perceptions affect later behaviors; cause precedes its effect. However, many researchers administer questionnaires and simultaneously collect data on past behaviors. The assumed effects follow the causes (e.g. Nicholson et al, 1976). Such research designs may underestimate the relation between questionnaire responses and behaviors (Morgan and Herman, 1976; Lawler, 1968; Wanous, 1974), as well as being illogical.

Horizontal Aggregation Across People

Another form of horizontal aggregation that can be done to increase the occurrences of the event is to increase N. As long as more individuals are sampled from the original population, the problems are not insurmountable. Basically, P remains unchanged (within limits of sampling fluctuation) with larger samples. The large N generates more occurrences of the event and, larger numbers of people in the sample have at least one occurrence. This larger sample provides more stable estimates of characteristics of those critical sample members.

Two apparent problems are time and money. Increasing N is expensive and potentially inefficient. So long as P is small, increasing N will also increase the number of people with scores of zero. If N is reasonably large to begin with, an investigator will already have a large sample of these subjects and stable estimates of their characteristics.

The danger of this practice, aside from its inefficiency, is that if populations are poorly specified, an investigator runs the risk of sampling members from different populations. Increases in N achieved at this cost will change the values of P , λ , and s^2 in unknown ways. Further, including members from populations with different probabilities of exhibiting the low

base rate variable results in obtained distributions of our dependent variables that are composites of the expected distributions in each population. These distributions are likely to be complex forms of binomial distributions (depending on the distribution of the P's from the multiple populations) (Parzen, 1960). Deviations of the obtained distribution from a theoretical random and independent distribution may be falsely interpreted as evidence for individual differences in proneness or liability for the event or behavior. True differences may actually lie in the environments of the populations. This interpretation may trigger a fruitless search for individual correlates or antecedents of the low base rate variable.

To reiterate, increasing N in order to increase the occurrence of the variable is, at best, inefficient. At worst it can be misleading if we sample individuals from multiple populations each with different values of P caused by environmental factors. The resulting composite distribution of X may misdirect research.

Vertical Aggregation into Groups

A third option that is exercised is to change slightly the definition of the infrequent event from a discrete, individual level variable indicating the frequency (including zero) of the behavior to a continuously distributed rate of occurrence characterizing groups. Rather than studying the absenteeism frequency or turnover by individuals, we study rates of absenteeism or turnover per work group, department, plant, organization, industry, nation, or any combination. This is the most frequent way economists study turnover--annual rates aggregated by organization or industry. Psychologists, on the other hand, have maintained an interest in the original dichotomous variable indicating whether or not an individual left a particular job or organization. Questions of ecological fallacy have

been dealt with elsewhere in this sourcebook (Glick and Roberts, 1980) as have the complex question of disentangling within group, between groups, and total effects and the numbers of associated degrees of freedom (Cronbach, 1976).

We note here that problems of levels of analysis as solutions to low base rates are no less serious because of their intractability. In fact, the very problems involved in choosing appropriate methods of analyzing effects of manipulations and conditions on individuals should sensitize us to the possible consequences of mistakes in analytic procedures. It is sufficient for the purposes of this chapter to note that group effects are not the same as individual effects, antecedents of rates of occurrence of a variable may be unrelated to antecedents of individual occurrences of that variable, and a complete explanation of a rate (in terms of variance explained) is analogous to an explanation of a between groups effect and says nothing about an individual level influence. As an example, we note that economic factors explain approximately 70% of the year to year variation in turnover rates aggregated at the level of the United States (Eagley, 1965). This says nothing about the theoretical maximum of the variance that can be explained by individual effects on individual turnover decisions made by members of one organization over a relatively short period of time. In fact, Hom, Katerberg, and Hulin (1979) have presented data showing that individuals' attitudes and behavioral intentions can explain approximately 70% of the variance of these individual decisions. Explanations of rates and individual occurrences are not competing for the same pool of variance.

Surrogate Variables

Another, more subtle, way to change the distribution of the event

without changing the dependent variable into a rate involves changing the value of P by changing the definition of the event we wish to study. Researchers frequently use surrogate variables, with less extreme base rates, for the original variable. We are not referring to the ubiquitous use of paper and pencil, self-report, recall measures of a behavior or event. The biases of retrospective measures are beyond the scope of this book. Assuming even a moderate amount of verisimilitude on the part of our anonymous respondents, severe base rate problems are expected in verbal reports mirroring the base rates of the original variables.

Consider the plight of an investigator who wants to study major, white-collar, theft in organizations. If an arbitrary value is selected, above which most would agree lies major theft and below which are amounts we would agree are minor, then confidential questionnaires asking for self-reports of past thefts of goods, time, or equipment are likely to yield very low base rates for major theft as so defined. However, investigators may include questionnaire items asking for reports of less extreme thefts down to paper clips and rubber bands. To circumvent the extreme base problem it is tempting to revise the definition of major theft until a point is reached that provides a more favorable base rate. Thus, an investigator might define theft as the taking of equipment in the amount of \$5,000 or more but regress down through smaller and smaller dollar amounts and finally analyze thefts of small amounts of supplies. Meals sent to wards for patients no longer there are consumed by nurses; aspirin disappears from open stock in infirmaries; paper and pencil costs in many organizations rise dramatically every year around the start of school; materials used in manufacturing processes are appropriated for home use; and workers frequently steal time by starting work late, taking breaks longer than the authorized period, and

quitting early. These are all thefts. But they are far from the original definition of the variable of interest. The vital question is the extent to which they represent occurrences of the same psychological construct. By using surrogates of the variable of interest, have we substituted variables that, in effect, tap different latent traits?

Standard solutions to questions of similarity of meaning and function of surrogates to the measures they stand for are unavailable because of the identical problems that led to the use of surrogates. With extreme base rates, relations between original measures and surrogates are severely restricted. The maximum relation between two variables with base rates of .01 and .60 is a phi coefficient of only .08. Standard correlational analysis will not yield evidence of substitutability. A common strategy for evaluating the substitutability of surrogate variables is to compute correlations among multiple operationalizations of that same trait. But the relations among variables thought to tap some underlying trait is not the question. The proper question has to do with the relation between each variable and the underlying trait or construct; a question not answered by usual convergent validity studies (Drasgow & Miller in press).

Consider again the example of white collar theft. If an investigator has available a large number of employees' responses to a lengthy questionnaire containing many items asking about thefts of various amounts and kinds, an analysis designed to reveal underlying latent trait might reveal the curves in Figure 1 showing relations between probabilities of reporting the indicated theft and the latent trait.

The ordinate is the probability of reporting engaging in the kind of theft described in an item conditioned on the amount of underlying trait. The abscissa is the amount of the underlying trait. Item 1 might ask

respondents if they had ever taken a piece of equipment valued at more than \$1000. Item 2 might ask about taking an item valued at more than \$100. Item 3 asks about taking typewriter ribbons, carbon paper, and ball point pens for personal use. Item 4 could ask about making Xerox copies of material for personal use.

The item characteristic curves show each item with a different base rate (referred to as item difficulty in the language of latent trait theory), different sensitivity for revealing small differences in the amount of the latent trait (discriminating power), and different relations to the underlying trait.

The important question about the item characteristic curves in Figure 1 is the implication of substituting items 3 and 4 for item 1 and 2 to define a group of employees who have engaged in theft. If such substitutions can be made, then an investigator can take advantage of the more favorable base rates of stealing supplies for personal use or using office equipment for personal reasons. If the price paid for the more favorable base rates is to change substantially the meaning of the criterion, then the costs may be too large.

Our purpose here is to point out the necessity of determining if the substituted measures are related to the latent trait in a manner even approximately similar to that of the original variable. Most frequently, isomorphism is assumed rather than demonstrated because standard techniques do not apply in such situations, because the analyses are mathematically complex and require large samples of subjects and items, and because, quite frankly, it is easier to assume something than to demonstrate it. The reader is referred to Lord and Novick (1968) or Warm (1978) for discussions of details of latent trait analyses.

Post Hoc Samples

Finally, we note in passing a "solution" to the problem of low base rate phenomena that seems to have little merit. This is the procedure of defining a sample, waiting for a period of time until some members of the sample display the behavior in question, drawing a matching sample of the same size from among those remaining individuals who do not display the behavior, and analyzing the combined post hoc samples as if they constituted a random sample from a population. Such procedures have been common in the analysis of suicides, dismissals from basic military training because of severe psychological disturbances, and juvenile delinquency. Examination of the procedures suggests that neither P nor N has been changed in the population, yet somehow, P is now .50 in the post hoc sample and the extreme base rate has disappeared. Conclusions drawn from analyses of antecedents of the behavior based on such samples are nearly always misleading, suggesting greater understanding of the problem than exists. We have observed $P(Y|X_1)$ --the probability of a person having an antecedent characteristic given that the person committed suicide, for example. The data that will be available when this information is used or in the analysis of those things truly antecedent--coming before in time--are of the form $P(X_1|Y)$ -- the probability of the critical behavior given the possession of Y . In our analysis we have conditioned on the wrong variable; we have conditioned on the critical variable or the consequent, and observed the antecedents rather than the reverse. Case studies of critical events or individuals frequently suffer from similar problems. The low base rate will not go away by constituting a post hoc sample with a base rate near .50 from a population with a base rate near .01. The illusion that we are dealing

with a frequent event and approximations of normal distributions will be revealed as statistical prestidigitation when the apparently impressive results are put to use.

Conclusions

Our discussion of the analysis of low base rate phenomena has led us through a number of common "solutions" to the analytic problems. Each solution creates problems of its own:

1. Increasing the time interval over which events are allowed to occur and data are aggregated will increase the number of low base rate events observed. This can lead to inconsistent results when researchers use different time intervals to assess and aggregate the events. Lack of attention to the time intervals different variables reflect can lead to inappropriate interpretation of these data. When a causal ordering among variables is presumed, the misinterpretations are confounded.

2. Increasing the sample size to increase the chance of observing infrequent events may produce a more heterogeneous sample (e.g. where data from different departments are combined). Greater sample heterogeneity can introduce environmental characteristics that correlate with the low base rate variable. Individual and environmental characteristics may be confounded by aggregation over large samples because people usually are not distributed randomly across environments. Unless the effects of both individual and environmental characteristics are assumed, there is no way to determine which set of factors is truly related to the phenomenon (Roberts, Hulin, and Rousseau, 1978).

3. Aggregating individual level data to the group level (e.g. turnover rate) to increase the variance in low base rate phenomena may alter the phenomena under study; factors influencing rates and individual occurrences

need not be the same.

4. Substituting low base rate variables with less extreme surrogates (e.g. substituting suicide attempts for actual suicides) may alter the trait underlying the event. Without comparing the characteristic curves of each variable, we cannot determine if the substitution is appropriate.

5. Using post-hoc samples or case studies where subjects are chosen after they manifest the behavior under study is likely to produce misleading results. Here researchers identify the consequence (the low base rate event) and then uncover its antecedents rather than the reverse, as in the case of postdiction rather than prediction. It is impossible to determine whether the "antecedents" are in any way causally related to the variable of interest.

We have explored some of the problems and pitfalls in the study of infrequently occurring events. It is apparent that researchers cannot afford to be cavalier with time, sampling, or levels of aggregation when infrequent events are studied. Moreover, researchers often simultaneously use several means of increasing the number of infrequent events observed; they compound the problems described above. All strategies described for dealing with low base rate problems create other problems. We can minimize our analytic difficulties by being aware of the distributional and conceptual issues in the study of infrequently occurring events

We recommend several steps:

(1) Researchers should specify clearly the time interval over which data are collected or aggregated and consider the different time intervals involved when comparing results of different studies or relating measures with different time referents.

(2) Characteristics of environmental settings that may affect the

occurrence of infrequent events should be considered when researchers attempt to observe more occurrences of low base rate phenomena by aggregating samples to increase total sample size. Assessments of environmental characteristics become increasingly important for interpreting data when sample heterogeneity increases.

(3) Studies of rates (individual occurrences aggregated to group levels) should be treated as distinct from studies of individual (unaggregated) occurrences.

(4) Latent trait, or a functionally equivalent, analysis should be employed when surrogates with less extreme base rates are used. Comparability of the meanings of the variables and latent traits that generate the distributions can be determined.

(5) Use of post-hoc samples and case studies should be limited to hypothesis generation; hypothesis testing by these methods is likely to be misleading.

Because infrequent events are often tantalizing or threatening, signalling system breakdowns, with large costs to society, we cannot afford to allow appropriate analysis of these events to be an infrequent event in itself.

References

- Cronbach, L.J. Research on classrooms and schools: Formulations of questions, design, and analysis. Stanford, CA.: Stanford Evaluation Consortium, Stanford University, 1976.
- Dragow, F. & Miller, H.E. Psychometric and Substantive issues in Scale Construction and Validation. Journal of Applied Psychology, in press.
- Eagly, R.V. Market power as an intervening mechanism in Phillips Curve analysis. Economics, 1965, 32, 48-64.
- Feller, W. An introduction to probability theory and its applications, Vol. I, New York: Wiley, 1950.
- Hom, P.W., Katerberg, R., & Hulin, C.L. Comparative examination of three approaches to the prediction of turnover. Journal of Applied Psychology, 1979, 64, 280-290.
- Kenny, D. A. Cross-lagged panel correlations: A test for spuriousness. Psychological Bulletin, 1975, 82, 887-903.
- Lawler, E.E. A Correlational-causal analysis of the relationship between expectancy attitudes and job performance. Journal of Applied Psychology, 1968, 52, 462-468.
- Lord, F.M. & Novick, M.R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.
- Miller, H.E., Katerberg, R., & Hulin, C.L. Evaluation of the Mobley, Horner, and Hollingsworth model of employee turnover. Journal of Applied Psychology, 1979, 64, 509-517.
- Morgan, L.G. & Herman, J.B. Perceived consequences of absenteeism. Journal of Applied Psychology, 1976, 61, 738-742.

Nicholson, N., Brown, C.A., & Chadwick-Jones, J.K. Absence from work and job satisfaction. Journal of Applied Psychology, 1976, 61, 728-737.

Parzen, E. Modern Probability theory and its applications. New York: Wiley, 1960.

Roberts, K.H., Hulin, C.L., & Rousseau, D.M. Developing an interdisciplinary science of organizations. San Francisco: Jossey-Bass, 1978.

Wanous, J.P. A Causal-Correlational Analysis of the job satisfaction and performance relationship. Journal of Applied Psychology, 1974, 59, 139-144.

Warm, T.A. A primer of item response theory. U.S. Coast Guard Institute, Oklahoma City: Technical Report #94, 1978.

Footnotes

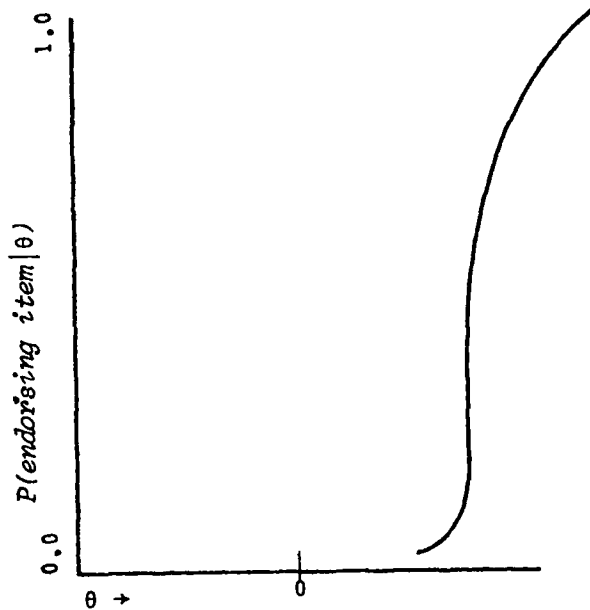
1. The authors would like to thank Howard Miller and John Komocar for their comments and suggestions on earlier drafts of this manuscript. Fritz Dragow was the source of many of our insights into the statistical complexities discussed. Portions of this work were supported by ONR Contract N000-14-75-C-0904, Charles L. Hulin, Principal Investigator.

Index

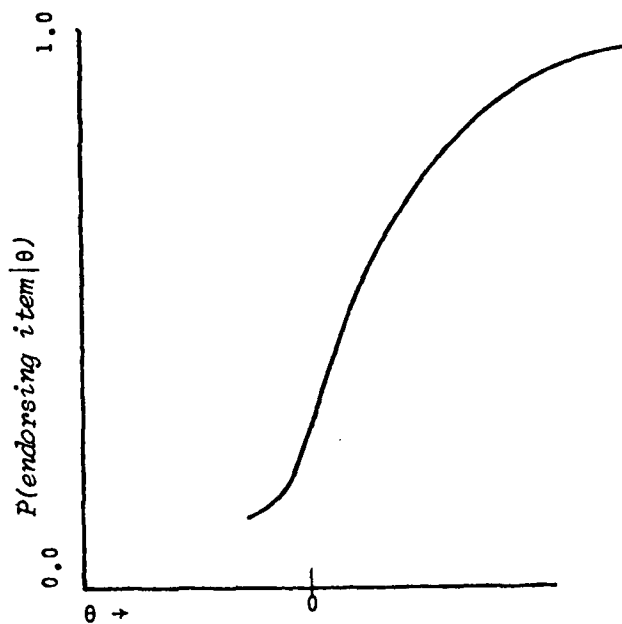
absence.....	6,7,20
absenteeism.....	1,5,6,8,10,19
accidents.....	1,3,4,8
aggregate.....	6,16
aggregated.....	5,6,7,10,11,16,18
aggregating.....	16,18
aggregation.....	5,8,9,10,16,17
aspirin.....	12
bacteria.....	3
base.....	1,2,3,5,7,10,11,12,13, 14,15,16,17,18
binomial.....	3,10
blizzard.....	5
coefficient.....	13
construct.....	1,13
convergent.....	13
correlation.....	6
cowboys.....	4
criterion.....	2
death.....	3
discontinuities.....	8
earthquake.....	5
ecological.....	10
economic.....	11
fluctuation.....	5,9
fluctuations.....	5
heterogeneity.....	16,18
horizontal.....	9
latent.....	13,14,18
maturation.....	5,7
null.....	3
organization.....	10,11
organizations.....	12,20
periodicity.....	7
poisson.....	3
pool.....	11
random.....	3,4,10
rodeo.....	4
skewed.....	3,5

steal.....12
stealing.....14
substitutability.....13
surrogate.....1,11,12,13
surrogates.....13,17,18
synchronicity.....5,8

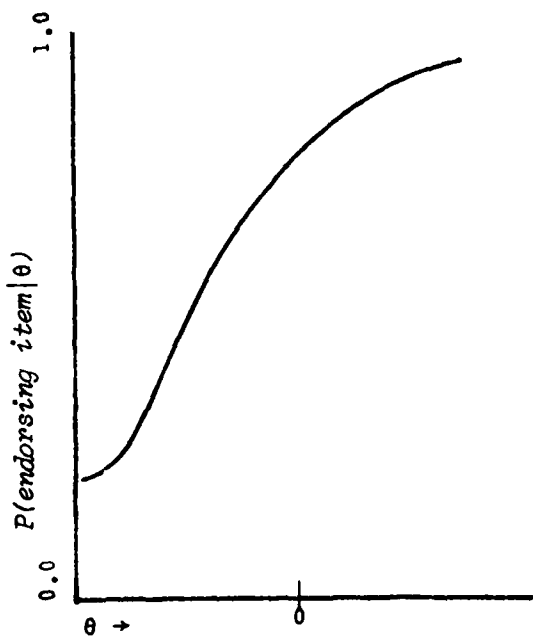
theft.....12,13,14
trait.....13,14,18



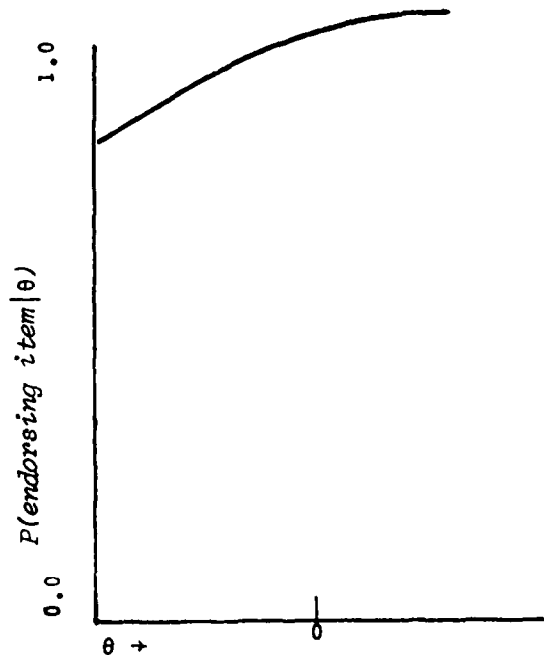
1a. Reported stealing equipment valued at more than \$1000.



1b. Reported stealing equipment valued at more than \$100.



1c. Reported stealing office supplies.



1d. Reported copying material for personal use.

Figure 1. Relationships between reports of stealing different material assessed by four hypothetical items and an underlying trait to steal, θ .

DISTRIBUTION LIST

Defense Documentation Center
ATTN: DDC-TC
Accessions Division
Cameron Station
Alexandria, VA 22314

Chief of Naval Research
Office of Naval Research
Code 452
800 N. Quincy Street
Arlington, VA 22217

Commanding Officer
ONR Branch Office
1030 E. Green Street
Pasadena, CA 91106

Commanding Officer
ONR Branch Office
536 S. Clark Street
Chicago, IL 60605

Commanding Officer
ONR Branch Office
Bldg. 114, Section D
666 Summer Street
Boston, MA 02210

Office of Naval Research
Director, Technology Programs
Code 200
800 N. Quincy Street
Arlington, VA 22217

Deputy Chief of Naval Operations
(Manpower, Personnel, and Training)
Director, Human Resource Management
Division (Op-15)
Department of the Navy
Washington, DC 20350

Deputy Chief of Naval Operations
(Manpower, Personnel, and Training)
Director, Human Resource Management
Plans and Policy Branch (Op-150)
Department of the Navy
Washington, DC 20350

Library of Congress
Science and Technology Division
Washington, DC 20540

Commanding Officer
Naval Research Laboratory
Code 2627
Washington, DC 20375

Psychologist
ONR Branch Office
1030 E. Green Street
Pasadena, CA 91106

Psychologist
ONR Branch Office
536 S. Clark Street
Chicago, IL 60605

Psychologist
ONR Branch Office
Bldg. 114, Section D
666 Summer Street
Boston, MA 02210

Deputy, Chief of Naval Operations
(Manpower, Personnel, and Training)
Scientific Advisor to DCNO (Op-01T)
2705 Arlington Annex
Washington, DC 20350

Deputy Chief of Naval Operations
(Manpower, Personnel, and Training)
Head, Research, Development, and
Studies Branch (Op-102)
1812 Arlington Annex
Washington, DC 20350

Chief of Naval Operations
Head, Manpower, Personnel, Training
and Reserves Team (Op-964D)
The Pentagon 4A578
Washington, DC 20350

Chief of Naval Operations
 Assistant, Personnel Logistics
 Planning (Op-987P10)
 The Pentagon, 5D772
 Washington, DC 20350

Naval Material Command
 Management Training Center
 NMAT 09M32
 Jefferson Plaza, BLDG #2, Rm 150
 1421 Jefferson Davis Highway
 Arlington, VA 20360

Navy Personnel R&D Center
 Washington Liaison Office
 Building 200, 2N
 Washington Navy Yard
 Washington, DC 20374

Commanding Officer
 Naval Submarine Medical
 Research Laboratory
 Naval Submarine Base
 New London, Box 900
 Groton, CT 06340

Naval Aerospace Medical
 Research Lab
 Naval Air Station
 Pensacola, FL 32508

National Naval Medical Center
 Psychology Department
 Bethesda, MD 20014

Naval Postgraduate School
 ATTN: Dr. Richard S. Elster
 Department of Administrative Sciences
 Monterey, CA 93940

Superintendent
 Naval Postgraduate School
 Code 1424
 Monterey, CA 93940

Officer in Charge
 Human Resource Management Detachment
 Naval Submarine Base New London
 P.O. Box 81
 Groton, CT 06340

Naval Material Command
 Program Administrator, Manpower,
 Personnel, and Training
 Code 08T244
 1044 Crystal Plaza #5
 Washington, DC 20360

Commanding Officer
 Naval Personnel R&D Center
 San Diego, CA 92152

Commanding Officer
 Naval Health Research Center
 San Diego, CA

Director, Medical Service Corps
 Bureau of Medicine and Surgery
 Code 23
 Department of the Navy
 Washington, DC 20372

CDR Robert Kennedy
 Officer in Charge
 Naval Aerospace Medical
 Research Laboratory Detachment
 Box 2940, Michoud Station
 New Orleans, LA 70129

Commanding Officer
 Navy Medical R&D Command
 Bethesda, MD 20014

Naval Postgraduate School
 ATTN: Professor John Senger
 Operations Research and
 Administrative Science
 Monterey, CA 93940

Officer in Charge
 Human Resource Management Detachment
 Naval Air Station
 Alameda, CA 94591

Officer in Charge
 Human Resource Management Division
 Naval Air Station
 Mayport, FL 32228

Commanding Officer
Human Resource Management Center
Pearl Harbor, HI 96860

Officer in Charge
Human Resource Management Detachment
Naval Base
Charleston, SC 29408

Human Resource Management School
Naval Air Station Memphis (96)
Millington, TN 38054

Commanding Officer
Human Resource Management Center
5621-23 Tidewater Drive
Norfolk, VA 23511

Officer in Charge
Human Resource Management Detachment
Naval Air Station Ehibbey Island
Oak Harbor, WA 98278

Commander in Chief
Human Resource Management Division
U.S. Naval Force Europe
FPO New York, 09510

Officer in Charge
Human Resource Management Detachment
COMNAVFOR JAPAN
FPO Seattle 98762

Chief of Naval Education
and Training (N-5)
ACOS Research and Program
Development
Naval Air Station
Pensacola, FL 32508

Navy Recruiting Command
Head, Research and Analysis Branch
Code 434, Room 8001
801 North Randolph Street
Arlington, VA 22203

Commander in Chief
Human Resource Management Division
U.S. Pacific Fleet
Pearl Harbor, HI 96860

Commanding Officer
Human Resource Management School
Naval Air Station Memphis
Millington, TN 38054

Commanding Officer
Human Resource Management Center
1300 Wilson Boulevard
Arlington, VA 22209

Commander in Chief
Human Resource Management Division
U.S. Atlantic Fleet
Norfolk, VA 23511

Commanding Officer
Human Resource Management Center
Box 23
FPO New York 09510

Officer in Charge
Human Resource Management Detachment
Box 60
FPO San Francisco 96651

Naval Amphibious School
Director, Human Resource
Training Department
Naval Amphibious Base
Little Creek
Norfolk, VA 23521

Naval Military Personnel Command
HRM Department (NMPC-6)
Washington, DC 20350

Chief of Navy Technical Training
ATTN: Dr. Norman Kerr, Code 0161
NAS Memphis (75)
Millington, TN 38054

Naval Training Analysis
and Evaluation Group
Orlando, FL 32813

Naval War College
Management Department
Newport, RI 02940

Headquarters, U.S. Marine Corps
ATTN: Dr. A.L. Slafkosky,
Code RD-1
Washington, DC 20380

Deputy Chief of Staff for
Personnel, Research Office
ATTN: DAPE-PBR
Washington, DC 20310

Army Research Institute
Field Unit - Leavenworth
P.O. Box 3122
Fort Leavenworth, KS 66127

Air University Library
LSE 76-443
Maxwell AFB, AL 36112

Air Force Institute of Technology
AFIT/LSGR (1st. Col. Umstot)
Wright-Patterson AFB
Dayton, OH 45433

AFMPC/DPMYP
(Research and Measurement Division)
Randolph AFB
Universal City, TX 78148

Dr. H. Russell Bernard
Department of Sociology
and Anthropology
West Virginia University
Morgantown, WV 26506

Dr. Michael Borus
Ohio State University
Columbus, OH 43210

Commanding Officer
Naval Training Equipment Center
Orlando, FL 32813

Commandant of the Marine Corps
Headquarters, U.S. Marine Corps
Code MPI-20
Washington, DC 20380

Army Research Institute
Field Unit - Monterey
P.O. Box 5787
Monterey, CA 93940

Headquarters, FORSCOM
ATTN: AFPR-HR
Ft. McPherson, GA 30330

Technical Director
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

AFOSR/NL (Dr. Fregly)
Building 410
Bolling AFB
Washington, DC 20332

Technical Director
AFHRL/ORS
Brooks AFB
San Antonio, TX 78235

Dr. Clayton P. Alderfer
School of Organization & Management
Yale University
New Haven, CT 06520

Dr. Arthur Blaiwes
Human Factors Laboratory, Code N-71
Naval Training Equipment Center
Orlando, FL 32813

Dr. Joseph V. Brady
The Johns Hopkins University
School of Medicine
Division of Behavioral Biology
Baltimore, MD 21205

Mr. Frank Clark
ADTECH/Advanced Technology, Inc.
7923 Jones Branch Drive, Suite 500
McLean, VA 22102

Mr. Gerald M. Croan
Westinghouse National Issues
Center
Suite 1111
2341 Jefferson Davis Highway
Arlington, VA 22202

Dr. John P. French, Jr
University of Michigan
Institute for Social Research
P.O. Box 1248
Ann Arbor, MI 48106

Dr. J. Richard Hackman
School of Organization
and Management
Yale University
56 Hillhouse Avenue
New Haven, CT 06520

Dr. Edna J. Hunter
United States International
University
School of Human Behavior
P.O. Box 26110
San Diego, CA 92126

Dr. Judi Komaki
Georgia Institute of Technology
Engineering Experiment Station
Atlanta, GA 30332

Dr. Edwin A. Locke
University of Maryland
College of Business and Management
and Department of Psychology
College Park, MD 20742

Dr. Richard T. Mowday
Graduate School of Management
and Business
University of Oregon
Eugene, OR 97403

Dr. Stuart W. Cook
University of Colorado
Institute of Behavioral Science
Boulder, CO 80309

Dr. Larry Cummings
University of Wisconsin-Madison
Graduate School of Business
Center for the Study of Organizational
Performance
1155 Observatory Drive
Madison, WI 53706

Dr. Paul S. Goodman
Graduate School of Industrial
Administration
Carnegie-Mellon University
Pittsburgh, PA 15213

Dr. Asa G. Hilliard, JR
The Urban Institute for
Human Services, Inc.
P.O. Box 15068
San Francisco, CA 94115

Dr. Rudi Klauss
Syracuse University
Public Administration Department
Maxwell School
Syracuse, NY 13210

Dr. Edward E. Lawler
Battelle Human Affairs
Research Centers
P.O. Box 5395
4000 N.E., 41st Street
Seattle, WA 98105

Dr. Ben Morgan
Performance Assessment
Laboratory
Old Dominion University
Norfolk, VA 23508

Dr. Joseph Olmstead
Human Resources Research
Organization
300 North Washington Street
Alexandria, VA 22314

Dr. Thomas M. Ostrom
The Ohio State University
Department of Psychology
116E Stadium
404C West 17th Avenue
Columbus, OH 43210

Dr. Irwin G. Sarason
University of Washington
Department of Psychology
Seattle, WA 98195

Dr. Saul B. Sells
Texas Christian University
Institute of Behavioral Research
Drawer C
Fort Worth, TX 76129

Dr. Richard Steers
Graduate School of Management
and Business
University of Oregon
Eugene, OR 97403

Dr. William H. Mobley
University of South Carolina
College of Business Administration
Columbia, SC 29208

Dr. Al. Rhode
Information Spectrum, Inc.
1745 S. Jefferson Davis Highway
Arlington, VA 22202

Dr. Donald Wise
MATHTECH, Inc.
P.O. Box 2392
Princeton, NJ 08540

Dr. George E. Rowland
Temple University, The Merit Center
Ritter Annex, 9th Floor
College of Education
Philadelphia, PA 19122

Dr. Benjamin Schneider
Michigan State University
East Lansing, MI 48824

Dr. H. Wallace Sinaiko
Program Director, Manpower Research
and Advisory Services
Smithsonian Institution
801 N. Pitt Street, Suite 120
Alexandria, VA 22314

Dr. Vincent Carroll
University of Pennsylvania
Wharton Applied Research Center
Philadelphia, PA 19104

Dr. Richard Morey
Duke University
Graduate School of Business
Administration
Durham, NC 27706

Dr. Lee Sechrest
Florida State University
Department of Psychology
Tallahassee, FL 32306